Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature

Abdullatif Köksal
Computer Engineering,
Boğaziçi University, Turkey
abdullatif.koksal@boun.edu.tr

Hilal Dönmez
Computer Engineering,
Boğaziçi University, Turkey
hilal.donmez@boun.edu.tr

Rıza Özçelik
Computer Engineering,
Boğaziçi University, Turkey
riza.ozcelik@boun.edu.tr

Elif Ozkirimli
Chemical Engineering,
Bogazici University, Turkey
Data and Analytics Chapter,
F. Hoffmann-La Roche AG, Switzerland
elif.ozkirimli@roche.com

Arzucan Özgür
Computer Engineering,
Boğaziçi University, Turkey
arzucan.ozgur@boun.edu.tr

Abstract

Coronavirus Disease of 2019 (COVID-19) created dire consequences globally and triggered an enormous scientific effort from different domains. Resulting publications formed a gigantic domain-specific collection of text in which finding studies on a biomolecule of interest is quite challenging for general purpose search engines due to terminology-rich characteristics of the publications. Here, we present Vapur, an online COVID-19 search engine specifically designed for finding related protein - chemical pairs. Vapur is empowered with a biochemically related entities-oriented inverted index in order to group studies relevant to a biomolecule with respect to its related entities. The inverted index of Vapur is automatically created with a BioNLP pipeline and integrated with an online user interface. The online interface is designed for the smooth traversal of the current literature and is publicly available at https://tabilab.cmpe.boun.edu.tr/vapur/.

1 Introduction

Coronavirus Disease of 2019 (COVID-19) outbreak had severe impacts on human health all around the world since December 2019, but also triggered an exceptional amount of scientific work. As of September 2020, PubMed recorded over 45K articles (Chen et al., 2020) related to COVID-19 since its inception, including works on diagnosis (Dorche et al., 2020), drug repurposing (Gao et al., 2020), and text-mining (Tarasova et al., 2020). As this tremendous literature keeps growing in the form of unstructured text, it also becomes more and more challenging for researchers to find the relevant information they need. Furthermore, the publications include named entities and chemical relations that challenge the general-purpose search engines. Therefore, it is of critical importance to build a search engine that can find relevant documents in this terminology-rich and domain-specific literature.

Biomedical named entity recognition (NER) and relation extraction can be utilized to semantically structure the publications around the biochemically related entities. When named entities and their relations are extracted, a document can be expressed as a set of triplets of the form $(Entity1, Entity2, Relation)$. This formulation can be converted to a semantic index from the related entities to the publications in order to enable retrieving relevant documents to a query by entity and relation matching. If the same entities are referenced with different words (e.g. ACE2, Angiotensin-converting enzyme 2, Q9BYF1), we can use named entity normalization to identify different mentions of the same entity.

In this work, we present Vapur, an online search engine to find related protein - compound pairs in the COVID-19 anthology. Vapur is empowered with a biochemical relation-based inverted index that is created through named entity recognition and relation extraction on CORD-19 abstracts (Wang et al., 2020a). In order to obtain the index, we first identify and normalize the named entities in the documents using a pre-trained model, BERN (Kim et al., 2019). Afterward, we determine if the entity pairs in the same sentence are related to each other by training a binary relation extraction model on the ChemProt data set (Krallinger et al., 2017) and obtain a list of related entities for each abstract. We utilize these sets to construct an inverted index that maps related entities to the documents they are mentioned.

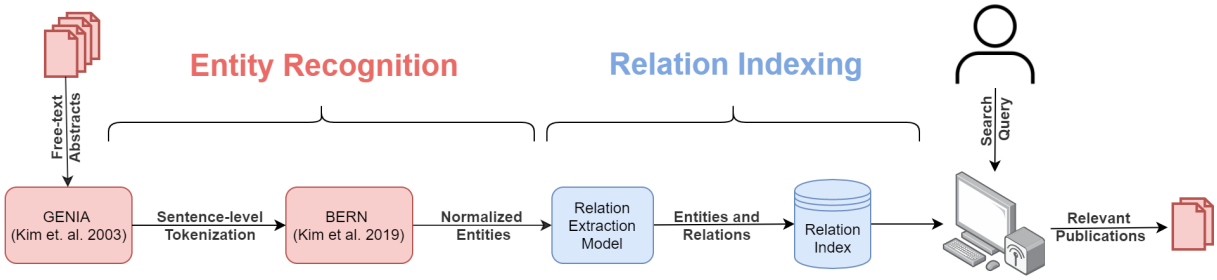

Figure 1: The workflow of the pipeline behind Vapur. We first split the abstracts to sentences and use BERN to detect the entities in the text. We then identify the biochemical relations with the relation extraction model that we trained and reform the output as an inverted index of relations. Vapur leverages this inverted index to retrieve relevant publications to the query as categorized by related entities.

We illustrate our pipeline in Figure 1 and present Vapur at `https://tabilab.cmpe.boun.edu.tr/vapur/`. We publicly share the code and models at `https://github.com/boun-tabi/vapur`.

2 Related Work

Extracting information from the growing body of biomedical literature has become even more important with the explosion of scientific research. Efforts such as ChemProt (Krallinger et al., 2017) aim to promote research in this direction by presenting PubMed abstracts in which proteins, chemicals, and their relations are manually labeled. Liu et al., 2018 and Lim and Kang, 2018 developed sequential models for chemical and protein relation extraction on ChemProt, while Peng et al., 2018 used an ensemble of deep models with SVM. However, all these models depend on the named entities to be recognized beforehand.

Named entity recognition is a widely studied topic in biomedical text mining. Conditional Random Fields (CRF) was a popular approach in early biomedical NER studies (Leaman et al., 2015; Wei et al., 2015, 2018) and with the rise of deep learning, sequential models were also integrated into the models (Sachan et al., 2018). Recently, transformers-based NER models were popularized, including BERN (Kim et al., 2019) which is a state-of-the-art biomedical named entity recognition and normalization tool that builds on BioBERT (Lee et al., 2020b) to identify and normalize the entities in a sentence. BERN is adopted by recent general-domain biomedical text mining studies (Jang et al., 2019) as well as a model to answer COVID-19 related questions based on CORD-19 (Lee et al., 2020a).

CORD-19 (Wang et al., 2020a) is a data set comprising scientific publications related to COVID-19 and provided a valuable resource for text mining studies. Su et al., 2020 and Esteva et al., 2020 developed question answering models with integrated information retrieval and summarization modules, while Wang et al., 2020b focused on extracting the chemical entities in CORD-19 and presented an automatically annotated NER data set, CORD-NER. The entities in CORD-NER are annotated with multiple tools, but they are not normalized. Here, for the first time, we apply named entity normalization and relation extraction to CORD-19 in order to create a biochemical relation-aware search engine.

3 Methods

3.1 Data sets

CORD-19 We create an entity-oriented search engine for CORD-19 abstracts. CORD-19 is a regularly updated data set that contains studies related to COVID-19 and SARS-CoV-2. We use the CORD-19 snapshot of August 23, 2020, that contains $\approx$ 233K documents for which $\approx$ 143K unique abstracts are provided. Vapur indexes only these abstracts, but it is able to return the linked full-paper, as well.

ChemProt We trained the binary relation extraction model using the ChemProt data set (Krallinger et al., 2017), a curation of PubMed abstracts whose entities and relations are manually annotated. ChemProt consists of three file types: abstracts, entities, and relations. An abstract file contains a PubMed abstract ID and its text, whereas an entity file lists the chemicals and proteins in the linked abstract. Last, a relation file reports the relation type between entity pairs in the same sentence whose relation types can be inferred from the sentence.

The relations in ChemProt are categorized into 11, where the first 10 categories (CPR:0 to CPR:9)

Relation ID	Description	Train Count	Dev Count	Test Count	Binary Label
CPR:0	*No description provided by ChemProt*	1	2	0	1
CPR:1	Part of	308	153	215	1
CPR:2	Regulator / Direct Regulator / Indirect regulator	1652	780	1743	1
CPR:3	Upregulator / Activator/ Indirect upregulator	777	552	667	1
CPR:4	Downregulator / Inhibitor / Indirect Downregulator	2260	1103	1667	1
CPR:5	Agonist / Agonist activator / Agonist inhibitor	173	116	198	1
CPR:6	Antagonist	235	199	293	1
CPR:7	Modulator / Modulator Activator / Modulator inhibitor	29	19	25	1
CPR:8	Cofactor	34	2	25	1
CPR:9	Substrate / Product of / Substrate product of	727	457	644	1
CPR:10	Not relation	241	175	267	0
Other	*No relation information between entities in this context*	11664	7780	9987	0

Table 1: Chemical - Protein Relations (CPR) in ChemProt. The entity pairs are annotated with 11 different relations in ChemProt and we refer to pairs whose relation information cannot be inferred from the context as *Other*. We report the instance count of each category and specify the label we assign to each category for binary relation extraction.

Statistic	Train	Dev	Test
# documents	1020	612	800
# sentences	18046	11294	15712
# entities	25752	15567	20828
# chemical mentions	13017	8004	10810
# unique chemical entities	3710	2517	3442
# protein mentions	12735	7563	10018
# unique protein entities	4610	3018	3757
# duplicate relations	98	86	158
# positive labels	6143	3339	5459
# negative labels	11903	7955	10253

Table 2: ChemProt summary statistics. We split abstracts to sentences via GENIA (Kim et al., 2003) and computed the statistics accordingly. We refer to relations in the same sentence with the same protein, chemical and relation type as duplicate relations and consider them only once during training. We observe that the # of mentions is considerably higher than the # of unique entities for both proteins and chemicals, emphasizing the need for normalization. In addition, negative entity pairs are significantly more frequent than the positive ones, but their distribution across the training, development, and test folds is similar.

denote various types of biochemical relations, while the last category (CPR:10) is reserved for "not relation" (i.e., the sentence explicitly states that there is no relation between the specified pair of entities). In this work, we focus on binary relation extraction, where we aim at determining whether a sentence states that there is a relation between the specified two entities or not. Therefore, we consider the entity pairs in the first 10 categories as positive samples. Likewise, since the sentences in the CPR:10 and the "Other" classes do not state that there is a relation between the corresponding pair of entities, we treat these two classes as the negative class. We describe each relation type and report their instance counts in the ChemProt data set in Table 1. We also provide summary statistics for ChemProt in Table 2.

3.2 Named Entity Recognition and Normalization

We used BERN (Kim et al., 2019), a neural NER architecture with an integrated normalizer that can perform in a multi-type setting, to identify and normalize the entities in CORD-19 abstracts. BERN is an ensemble of existing tools and models (Wei et al., 2015; Leaman et al., 2015; D'Souza and Ng, 2015; Wei et al., 2016, 2018; Lee et al., 2020b) and outputs a set of IDs for each recognized entity, given a sentence.

In this study, we first tokenized the abstracts with GENIA (Kim et al., 2003) and identified 1.23M

Raw Sentence	EGFR inhibitors currently under investigation include the small molecules gefitinib and erlotinib.
Preprocessed Form I	<e2>EGFR</e2> inhibitors currently under investigation include the small molecules <e1>gefitinib</e1> and erlotinib.
Preprocessed Form II	<e2>EGFR</e2> inhibitors currently under investigation include the small molecules gefitinib and <e1>erlotinib</e1>.

Table 3: An example of preprocessing. The raw sentence contains two chemicals (*gefitinib*, *erlotinib*) and one protein (EGFR), creating two protein - chemical pairs. Thus, we create two different forms of the sentence to encode each protein - chemical pair separately. We use <e1> and <e2> tags to enclose chemicals and proteins, respectively.

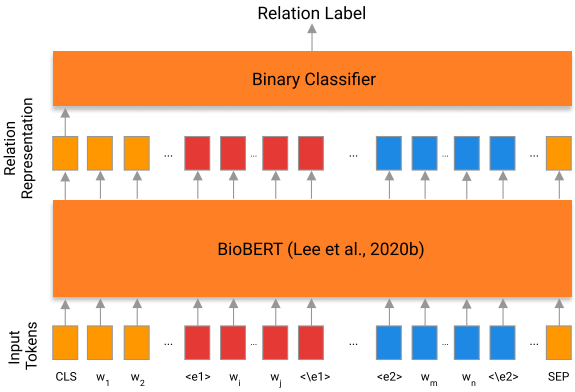

Figure 2: Relation Extraction Model. Our model is composed of a BioBERT and a binary classification layer. We tokenize the input sentences by BioBERT tokenizer and enclose the biochemical entities with special tags (<e1> and <e2>). BioBERT also adds *CLS* and *SEP* tags for fixed length-relation representation and sentence separation, respectively. We use BioBERT output of *CLS* token to train our binary classifier.

sentences. We used BERN to extract named entities and their IDs in these sentences and recognized 1.58M entities in total, in which 171K are chemicals and 318K are proteins. With normalization, we computed the number of unique chemicals and proteins as 20K and 70K, respectively.

3.3 Relation Extraction

We propose a BioBERT-based (Lee et al., 2020b) model to identify related entities in CORD-19 abstracts. To this end, we first preprocessed the sentences to explicitly encode the entities in the input and then fine-tuned BioBERT with the preprocessed sentences in ChemProt. We used the binary labels in Section 3.1 as outputs.

Preprocessing: We surrounded the named entities in the sentences with opening and closing tags to mark their location. We tagged chemicals with <e1> and </e1> and proteins with <e2> and </e2> to encode entity type and location during training. When a sentence had multiple chemical - protein pairs, we considered each pair separately and created copies of the sentence with different tags. Table 3 provides an example of input and its two preprocessed forms.

Model: We used BioBERT to create a binary model that decides if two entities in the same sentence are related. We added a single-layer binary log-softmax classifier to BioBERT and trained the classifier with the inputs and outputs from the previous step. BioBERT creates a fixed-length relation representation for the starting tag *CLS* and we used this representation to vectorize the sentences. We trained the binary classifier with cross-entropy loss where we considered the pairs in CPR:1-9 as positive and pairs in CPR:10 or no reported relation (i.e., the "Other" Type in Table 1) as negative. Figure 2 illustrates our relation extraction model.

Experimental setup: We conducted a hyperparameter search with different optimizers, learning rates, and weight decays. We trained a model 10 times per parameter combination and selected the best one with respect to F-score on the development set. We found that the AdamW optimizer with a 0.00003 learning rate and 0.1 weight decay yields the best results. We selected the best setting for relation extraction and computed mean precision, recall, and F1-score as well as the standard deviation on the ChemProt development (dev) and test sets.

4 Vapur

Vapur is an online search engine with a focus on finding related proteins and chemicals from CORD-19, a compilation of COVID-19 literature (Wang et al., 2020a). Vapur uses an inverted index that is constructed through mining the relations and entities in CORD-19 and is able to retrieve relevant

Figure 3: A search scenario on Vapur. When *Interleukin 1b* is entered as the query, Vapur lists the biochemically related compounds and displays the publications that mention the relation. Unique IDs of each chemical entity are displayed to help the user for further research. Similar proteins to *Interleukin 1b* are also listed in order to let the user explore the index smoothly.

documents to the queried biomolecule as categorized by the relation extraction model. Vapur's inverted index explicitly represents relations as related entity pairs and maps each relation to the documents in which the relation was mentioned.

We also interpret the relation-oriented index of Vapur as a graph to infer relations that cannot be inferred from the inverted index trivially. We construct a graph from the index such that the nodes represent biomolecules and the edges denote biochemical relations. The resulting graph satisfies the bipartiteness property, since we identify only the relations between proteins and chemicals during binary relation extraction. Table 4 demonstrates the summary statistics of this bipartite graph computed via networkx (Hagberg et al., 2008).

We leverage the bipartite graph structure to identify similar biomolecules of the same type. We use SimRank (Jeh and Widom, 2002) to compute pairwise node similarity and list the five most similar entities to the query. The aim is to encourage the user to explore the information in CORD-19 not only for exact matches to the query, but also similar entities.

The entities are represented as equivalence classes learned from named entity recognition and

Graph Property	Statistic
# Nodes (Entities)	12384
# Chemical Nodes	5018
# Protein Nodes	7366
# Edges (Relations)	17657
# Connected Components	807
# Nodes in the Largest Component	10194
# Edges in the Largest Component	16226
# Diameter of the Largest Component	18

Table 4: Summary statistics for the bipartite graph of the inverted index. We observe that number of chemicals and proteins are comparable to each other and the number of relations is close to the number of nodes, indicating the sparsity. In addition, the graph is formed of 807 connected components but the largest one is significantly larger than the others.

normalization steps. For instance, "Interleukin-1b" is represented as {interleukin-1b, IL-1beta, IL1B, HGNC:5992, ..., BERN:323737602} in the index and Vapur retrieves related entities and relevant documents with a mention "Interleukin-1b" even if the query is "IL1B". The equivalence classes cover a wide range of mention types from free-text to BERN IDs to present a flexible search experience in terms of search terms.

Vapur represents each mention type as a string and adopts a 3-gram based matching algorithm to search the queried entity in its index. Given a query, Vapur first creates a multi-set of all 3-grams of the query and computes the similarity of this set to all 3-gram multi-sets of the mentions in the index, which are pre-computed. Vapur measures the similarity of the query to a mention by generalized Jaccard similarity:

$$J(Q, M) = \frac{\sum_i min(Q_i, M_i)}{\sum_i max(Q_i, M_i)}$$

where Q and M are query and mention vectors that store the count of each 3-gram. Generalized Jaccard similarity makes Vapur more robust while retrieving results for a mistyped query.

When the related entities are found and their related abstracts are retrieved, Vapur ranks the related entities by the number of times they are co-mentioned with the queried entity. Each entity is displayed with its most frequent mention alongside the links to the papers. Figure 3 illustrates an example search scenario on Vapur, which is publicly available at https://tabilab.cmpe.boun.edu.tr/vapur/

5 Results and Discussion

In this work, we built Vapur, a CORD-19 search engine empowered by text-mining. We applied named entity recognition, named entity normalization, and relation extraction to CORD-19 abstracts and automatically created an inverted index from related entities to publications. Therefore, the performance of the named entity extraction and relation extraction modules is critical to identify relevant works. Since BERN was already evaluated for named entity recognition and normalization on the biomedical domain (Kim et al., 2019), here we focus on evaluating the relation extraction model.

We built the relation extraction model by fine-tuning a binary classifier on BioBERT and computed precision, recall, and F1-Score on the dev and test sets of ChemProt. We also fine-tuned the same classifier using English BERT based model and report the results in Table 5.

Fold	Model	Precision	Recall	F1-Score
Dev	BERT	0.718±0.027	0.737±0.026	0.727±0.010
	BioBERT	**0.742±0.036**	**0.829±0.035**	**0.782±0.012**
Test	BERT	0.759±0.023	0.710±0.026	0.733±0.007
	BioBERT	**0.791±0.026**	**0.766±0.036**	**0.777±0.010**

Table 5: BERT- and BioBERT-based fine-tuning results for binary relation extraction. We computed precision, recall, and F1 on the dev and test folds of ChemProt. The results show that BioBERT-based model outperforms BERT-based model in terms of all metrics on both folds.

Relation ID	Test Set Accuracy
CPR:1	0.661 ± 0.053
CPR:2	0.667 ± 0.043
CPR:3	0.855 ± 0.015
CPR:4	0.874 ± 0.027
CPR:5	0.850 ± 0.045
CPR:6	0.856 ± 0.050
CPR:7	0.890 ± 0.044
CPR:8	0.550 ± 0.066
CPR:9	0.617 ± 0.060
CPR:10	0.828 ± 0.053
Other	0.892 ± 0.020

Table 6: Test performance of the relation extraction model by CPR label. We used the best model trained with 10 random seeds to predict test set pairs and report mean and standard deviation of the test accuracy by label. The results display that the relation extraction model is the most successful at the *Other* category, indicating its high context-awareness.

Table 5 demonstrates that BioBERT obtained higher scores in terms of all three metrics on both folds, indicating that BioBERT is superior to BERT as the pre-trained language model for relation extraction on ChemProt. We relate this with the fact that BioBERT is trained with a more domain-related text and BERT observed texts from a wider range.

We further investigated the performance of our relation extraction model by computing test accuracy for each chemical - protein relation (CPR) category and pairs with no assigned CPR label (i.e., Other). We observe that the relation extraction model achieved the highest accuracy in the *Other*

Sentence in CORD-19	Incorrectly Labeled Entity
<e1>Alanine</e1> substitution of either Arg-76 or Tyr-94 in the N-terminal domain of <e2>IBV N protein</e2> led to a significant decrease in its RNA-binding activity and a total loss of the infectivity of the viral RNA to Vero cells.	**Alanine**
<e2>Rat microsomal aldehyde dehydrogenase</e2> (msALDH) has no amino-terminal signal sequence, but instead it has a characteristic hydrophobic domain at the <e1>carboxyl</e1> terminus (Miyauehi, K., R.	**carboxyl**
Also, the <e2>protease Factor Xa</e2>, a target of <e1>Ben</e1>-HCl abundantly expressed in infected cells, was able to cleave the recombinant and pseudoviral S protein into S1 and S2 subunits, and the cleavage was inhibited by Ben-HCl.	**Ben**

Table 7: Sample sentences with incorrectly labeled entities. The entity types of *Alanine, carboxyl*, and *Ben* are incorrectly predicted as chemical by BERN during the named entity recognition step. Consequently, the relation extraction model incorrectly identified a biochemical relation.

category, indicating that the model can successfully identify whether the context is sufficient to deduce a relation between the entities or not. We relate this with using BioBERT, which is a contextual model for relation representation. Context-awareness enables Vapur to eliminate the documents that mention the queried entity in irrelevant contexts and to retrieve a document only if it contains relation information for the query.

We analyzed 41 sample sentences that we predicted to be positive to discover the limitations of our methodology. Most of the incorrect relation labels are due to incorrect entity assignment by BERN. In some cases, parts of the protein sequence such as N-terminal, carboxyl terminal or residue names such as Asp238 are recognized as compound. Table 7 illustrates sample sentences with incorrectly labeled entities. More examples that are manually checked by a domain expert are presented on our GitHub page as an Appendix in the Readme file.

6 Conclusion

We presented Vapur, a search engine with an emphasis on identifying protein - chemical relations in the COVID-19 domain. Vapur is the first search engine we know of that uses relation extraction to construct an inverted index of related biochemical entities. Thanks to the relation extraction model, Vapur can categorize relevant documents to a query with respect to related chemical entities and present organized search results.

We evaluated the relation extraction model on ChemProt and observed that BioBERT outperforms English BERT on both development and test sets in terms of precision, recall, and F1-score. We further analyzed the model predictions and demonstrated the high specificity of the model at identifying contexts with biochemical relation information.

Although we focused on COVID-19 in this work, our approach can find application in any biomedical domain with a different choice of database to index. We believe that domain-specific scientific search engines will gain more interest in the future since COVID-19 is unlikely to be the last global health crisis for human-kind. In order too promote the development and use of search engines to help coping with such difficult times, we make the underlying pipeline and Vapur available at `https://github.com/boun-tabi/vapur` and `https://tabilab.cmpe.boun.edu.tr/vapur/`, respectively.

Acknowledgments

TUBITAK-BIDEB 2211-A Scholarship Program (to A.K. and R.O.), and TUBA-GEBIP Award of the Turkish Science Academy (to A.O.) are gratefully acknowledged. E.O. is an employee of F. Hoffman - La Roche AG, Switzerland.

References

Q. Chen, A. Allot, and Z. Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193.

Maryam Sharifian Dorche, Philippe Huot, Micheal Osherov, Dingke Wen, Alexander Saveriano, Paul Giacomini, Jack P Antel, and Ashkan Mowla. 2020. Neurological complications of coronavirus infection; a comparative review and lessons learned during the covid-19 pandemic. *Journal of the Neurological Sciences*, page 117085.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based

entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.

Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv preprint arXiv:2006.09595*.

Kaifu Gao, Duc Duy Nguyen, Jiahui Chen, Rui Wang, and Guo-Wei Wei. 2020. Repositioning of 8565 existing drugs for covid-19. *arXiv preprint arXiv:2005.10028*.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Ye Eun Jang, Insu Jang, Sunkyu Kim, Subin Cho, Daehan Kim, Keonwoo Kim, Jaewon Kim, Jimin Hwang, Sangok Kim, Jaesang Kim, Jaewoo Kang, Byungwook Lee, and Sanghyuk Lee. 2019. ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Research*, 48(D1):D817–D824.

Glen Jeh and Jennifer Widom. 2002. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543.

Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, GP Rodríguez, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(S1):S3.

Jinhyuk Lee, Sean S Yi, Minbyul Jeong, Mujeen Sung, Wonjin Yoon, Yonghwa Choi, Miyoung Ko, and Jaewoo Kang. 2020a. Answering questions on covid-19 in real-time. *arXiv preprint arXiv:2006.15830*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020b. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sangrak Lim and Jaewoo Kang. 2018. Chemical–gene relation extraction using recursive neural network. *Database*, 2018. Bay060.

Sijia Liu, Feichen Shen, Ravikumar Komandur Elayavilli, Yanshan Wang, Majid Rastegar-Mojarad, Vipin Chaudhary, and Hongfang Liu. 2018. Extracting chemical–protein relations using attention-based neural networks. *Database*, 2018. Bay102.

Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Chemical-protein relation extraction with ensembles of svm, cnn, and rnn models. *arXiv preprint arXiv:1802.01255*.

Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P Xing. 2018. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine Learning for Healthcare Conference*, pages 383–402.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. 2020. Caire-covid: A question answering and multi-document summarization system for covid-19 research. *arXiv preprint arXiv:2005.03975*.

Olga Tarasova, Sergey Ivanov, Dmitry A Filimonov, and Vladimir Poroikov. 2020. Data and text mining help identify key proteins involved in the molecular mechanisms shared by sars-cov-2 and hiv-1. *Molecules*, 25(12):2944.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020a. Cord-19: The covid-19 open research dataset. *ArXiv*.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020b. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*.

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2016. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, 32(12):1907–1910.

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, and Zhiyong Lu. 2018. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87.